

A data driven trimming procedure for robust classification

Marina Agulló Antolín*, Eustasio del Barrio* and Jean-Michel Loubes†

January 19, 2017

Abstract

Classification rules can be severely affected by the presence of disturbing observations in the training sample. Looking for an optimal classifier with such data may lead to unnecessarily complex rules. So, simpler effective classification rules could be achieved if we relax the goal of fitting a good rule for the whole training sample but only consider a fraction of the data. In this paper we introduce a new method based on trimming to produce classification rules with guaranteed performance on a significant fraction of the data. In particular, we provide an automatic way of determining the right trimming proportion and obtain in this setting oracle bounds for the classification error on the new data set.

AMS subject classifications: Primary, 62H10; secondary, 62E20

Keywords: classification, outliers, robust statistics, trimming procedure.

1 Introduction

In the usual classification setting we observe a collection of pairs of i.i.d copies $(Y_i, X_i) \in \{0, 1\} \times \mathbb{R}^p$ with $i = 1, \dots, n$ of a random variable (Y, X) with distribution P . Y is the label to be forecast according to the value of the variables X . A classifier is a function $g : \mathbb{R}^p \mapsto \{0, 1\}$ that predicts the label of an observation. An observation is misclassified if $Y \neq g(X)$. Hence, the performance of a classifier can be measured by its classification error defined as $R(g) = P((y, x) \in \{0, 1\} \times \mathbb{R}^p : y \neq g(x))$. During the last decades, the classification problem a.k.a pattern recognition has been extensively studied and there exists a large variety of methods to find optimal classifiers in different settings. We refer for instance to [16], [7] or [10] and references therein for a survey.

When the number of observations grows large or in a high dimensional case, some of the data may contain observation errors and may be considered as contaminating data.

*IMUVa. Universidad de Valladolid. 7, paseo de Belén, 47011 Valladolid. Spain. These authors have been partially supported by the Spanish Ministerio de Economía y Competitividad, grants MTM2014-56235-C2-1-P, and MTM2014-56235-C2-2, and by Consejería de Educación de la Junta de Castilla y León, grant VA212U13.

†Institut de Mathématiques de Toulouse, Université Paul Sabatier. 118, route de Narbonne F-31062 Toulouse Cedex 9, loubes@math.univ-toulouse.fr

The presence of such observations, if not removed, hampers the efficiency of classifiers since many classification methods are very sensitive to outliers. Actually, if the learning set is too corrupted, training a classifier over this set leads to bad classification rates. Hence there is a growing need for robust methods to tackle such issue. Pioneered in [18], we refer to [15] for a review of robust methods.

A solution to cope with this issue is to allow the classifier not to label all points but to reject some observations that may seem too difficult to be classified. This point of view is studied in [17] and [4]. Another general idea is to remove a proportion of contaminating data to guarantee the robustness of the method. Such data are defined as outliers in the sense that they are far from the model used to generate the data. Yet detecting automatically outliers is a difficult task since their mere definition is unclear and highly depends on each particular case. Much research has been done in this direction and many analysis provide several ways of determining whether an observation is an outlier. For instance in [9], in the case of SVM classification, the author proposes to remove observations using an outlier map. In [6], the authors rely on a function that measures the impact of contamination of the distribution on the classification error obtained by minimizing the empirical risk criterion. In a regression framework, Lasso estimators suffer from outliers. They can also be modified in order to enhance robustness as in [5], [22] or in [1] where the authors discard the points for which the residuals are the largest.

In a probabilistic framework, removing observations that achieve bad classification error, corresponds to trimming the initial distribution of the observations and replacing it by a similar distribution Q up to some data that will be considered as outliers for the classification rule. Trimming methods for data analysis have been described firstly in [21] and later some statistical properties are studied in [13] or in [8]. Yet very few theoretical results exist to study how to choose the actual boundary between an acceptable observation and an outlier. Moreover, in that case little is known about whether this choice modifies the classification error or how. For both theoretical and practical purposes, such a choice must be guided in order to take into account the amount of variability generated by the corrupted data.

In this paper we provide some theoretical guarantee to choose the level of data to be removed. For this we consider the set of trimmed distributions obtained from the initial distribution of the data and look for an automatic rule that reduces the classification error of a collection of classifiers by removing some properly selected observations. The more data is removed, the easier it becomes to classify the data, leading to a perfect classification if the classification rate is small enough. Yet, removing too many data reduces the interest of the classification procedure. If too many observations are left aside then the chosen classification rule may be good for distribution which is possibly very far from the true distribution of the data. We provide in this work an empirical rule that automatically selects the minimum level of trimming to reduce the classification error for a class of classifiers. Simultaneously, the best classifier for the trimmed set of observations is chosen among a collection of classification rules and for this, we prove an oracle inequality that governs the statistical properties of this methodology.

The paper falls into the following parts. Section 2 is devoted to the description of the probabilistic framework of outlier selection using trimming distributions. We precisely define the trimmed classification errors and their relationships with the usual classification

errors for both the empirical and theoretical error. In section 3 we provide the automatic selection rule for the trimming level and the best trimmed classifier for which we provide and oracle inequality. This model selection result is illustrated with the case of linear classifiers. Section 4 provides some conclusions and perspectives for these results. The proofs and some technical results are gathered in the Appendix.

2 Partial Classification with trimming

As in the introduction, we assume that we observe an n i.i.d sample $(Y_i, X_i)_{i=1, \dots, n} \in \{0, 1\} \times \mathbb{R}^p$ with distribution P . Set $g : \mathbb{R}^p \mapsto \{0, 1\}$, a classification rule, we denote the classification error as

$$R(g) = P((y, x) \in \{0, 1\} \times \mathbb{R}^p : y \neq g(x)).$$

Since the underlying distribution of the observations is unknown, the classification error R is estimated by its empirical counterpart, the empirical error defined as

$$R_n(g) := \frac{1}{n} \sum_{i=1}^n I_{(g(X_i) \neq Y_i)},$$

where $I_{(g(X) \neq Y)} = 1$ if $g(X) \neq Y$ and 0 otherwise.

Trimming a data sample of size n is usually defined as discarding a given fraction of the data while reweighing the other part. Let α be the proportion of observations we can trim, and consider that $n\alpha = k \in \mathbb{N}$. Then, trimming consists of removing k observations and giving weight $1/(n - k)$ to the rest. Among all the possible trimmings, we will call *empirical trimmed classification error* the one that minimizes the sum

$$R_{n,\alpha}(g) := \min_{w \in W} \sum_{j=1}^n w_j I_{(g(x_j) \neq y_j)} \quad (1)$$

with

$$W = \{w = (w_1, \dots, w_n) / 0 \leq w_i \leq \frac{1}{n(1 - \alpha)}; i = 1, \dots, n \wedge \sum_{i=1}^n w_i = 1\}. \quad (2)$$

To study the theoretical counterpart of this quantity, which we will call trimmed classification error, we will consider the set of trimmed distributions as follows. From a probabilistic point of view, trimming a distribution consists of replacing the initial distribution of the observations by a new measure built by a partial removal of points in the support of the initial distribution. We thus can provide the following definition for the trimming of a distribution. Here, \mathcal{P} denotes the set of probabilities on $\{0, 1\} \times \mathbb{R}^p$.

Definition 2.1. Given $\alpha \in (0, 1)$, we define the set of α -trimmed versions of P by

$$\mathcal{R}_\alpha(P) := \left\{ Q \in \mathcal{P} : Q \ll P, \frac{dQ}{dP} \leq \frac{1}{1 - \alpha} P - a.s. \right\}.$$

This entails that a trimmed distribution $Q \in \mathcal{R}_\alpha(P)$ can be seen as a close modification of a distribution P obtained by removing a certain quantity of data (see [2] and [3]). When dealing with a classification rule, one is interested in looking for the data for which the classification rule performs well. Hence, we aim at improving the classification error by changing the underlying distribution of the observations using a trimming scheme in order to modify the distribution but yet in a controlled, limited way. With this goal we introduce the trimmed classification error for a rule $g : \mathbb{R}^p \mapsto \{0, 1\}$.

Definition 2.2. Given $\alpha \in (0, 1)$, we define the *trimmed classification error* of a rule as the infimum of the α -trimmed probabilities of misclassifying future observations

$$R_\alpha(g) := \inf_{Q \in \mathcal{R}_\alpha(P)} Q(g(x) \neq y).$$

There is a simple relation between the trimmed classification error and the general classification error as the next result shows.

Proposition 2.1. *Given a trimming level $\alpha \in (0, 1)$ and a classification rule g ,*

$$R_\alpha(g) = \frac{1}{1 - \alpha} (R(g) - \alpha)_+. \quad (3)$$

This proposition shows the effect of trimming on classification. We write g_B for the Bayes classifier, namely, the classification rule that yields the minimal classification error. We also write

$$Err(P) := \min_g R(g) = R(g_B),$$

for the *Bayes classification error*. The optimal trimming for classification removes the misclassified points in such a way that if the classification error is less than the percentage of points that can be removed, then all the points are classified without error.

Similar to the Bayes rule, we can define a trimmed Bayes classification rule and the trimmed Bayes error as follows.

Definition 2.3. An α -trimmed Bayes classifier or α -trimmed Bayes classification rule is a classifier that achieves the minimum α -Trimmed classification error

$$g_B^\alpha := \arg \min_g R_\alpha(g).$$

The corresponding classification error is thus the α -Trimmed Bayes error defined as

$$Err_\alpha(P) := \inf_{Q \in \mathcal{R}_\alpha(P)} Err(Q) = \min_g R_\alpha(g) = R_\alpha(g_B^\alpha).$$

The following proposition compares these two errors.

Proposition 2.2.

$$Err_\alpha(P) = \frac{(R(g_B) - \alpha)_+}{1 - \alpha} = \frac{(Err(P) - \alpha)_+}{1 - \alpha}.$$

If $\text{Err}(P) \leq \alpha$ then $\text{Err}_\alpha(P) = 0$, but if $\text{Err}(P) > \alpha$ then $\text{Err}_\alpha(P) = \frac{(\text{Err}(P) - \alpha)_+}{1 - \alpha} > 0$, which indicates that $\text{Err}_\alpha(P) = 0$ is equivalent to $\text{Err}(P) \leq \alpha$. This means, the minimum α which gives us the perfect separation is the value that corresponds to the Bayes error.

Usually we do not look for the optimum classifier among all possible classifiers, but we restrict ourselves to a smaller class of classifiers. Let \mathcal{F} be a class of classifiers and let $f^* \in \mathcal{F}$ be the classifier which gives us the minimum classification error within the class. We denote as $R(\mathcal{F})$ the minimum classification error in \mathcal{F} , that is

$$R(\mathcal{F}) := \min_{f \in \mathcal{F}} R(f) = R(f^*).$$

In the same way we denote the trimmed error of the class \mathcal{F} as $R_\alpha(\mathcal{F})$. Hence, given Proposition 2.1,

$$R_\alpha(\mathcal{F}) := \min_{f \in \mathcal{F}} R_\alpha(f) = \min_{f \in \mathcal{F}} \frac{(R(f) - \alpha)_+}{1 - \alpha}.$$

The classifier that minimizes $R(f)$ also minimizes this quantity and so the classifier that minimizes the error in the class \mathcal{F} is a minimizer of the trimmed error in the class.

Proposition 2.1 can be trivially applied to the empirical trimmed classification error introduced in (1). For convenience we state this fact in the following result.

$$R_{n,\alpha}(g) := \inf_{Q \in \mathcal{R}_\alpha(P_n)} Q(g(X) \neq Y),$$

where P_n is the empirical distribution of P .

Corollary 2.3. *Let g be a given classifier, α a fixed trimming level and $n \in \mathbb{N}$ the sample size,*

$$R_{n,\alpha}(g) = \frac{1}{1 - \alpha} (R_n(g) - \alpha)_+. \quad (4)$$

In empirical risk minimization methods (see for instance [19] and references therein), the empirical classification error $R_n(g)$ is used as an estimator of $R(g)$. Among other good properties, $R_n(g)$ is unbiased as an estimator of $R(g)$. This does not hold for trimmed errors. The following proposition provides a control over this quantity and shows that the empirical classification is still an asymptotically unbiased estimate of the classification error.

Proposition 2.4. *For a given trimming level α and a given classifier g*

$$0 \leq E(R_{n,\alpha}(g)) - R_\alpha(g) \leq \frac{\sqrt{R(g)}}{\sqrt{2n}(1 - \alpha)}.$$

3 Optimal selection of trimming levels in classification

3.1 Main Result

Trimmed models enable to decrease the classification error in such a way that the loss of information of using less observations can be quantified and controlled. As in any robust procedure, we aim at selecting the amount of data to be removed, which, in this setting,

corresponds to the optimal trimming level. Actually the aim is to find a data driven $\hat{\alpha}$ such that the classification risk is minimized without removing a too large quantity of information about the initial distribution. We know that the bigger the trimming is, the smaller the error will be, but the more data we trim, the less information our model will keep. To look for an equilibrium we will introduce a penalization which will depend on the size of the chosen trimming level. For the sake of clarity we present first an oracle bound in the toy setup in which we only consider a fixed classification rule and we aim at choosing the right trimming proportion. Later we present a more general result which will deal with the more realistic case in which the classifier is chosen within a more general collection of models.

Theorem 3.1. *Let $\xi_1 = (Y_1, X_1), \dots, \xi_n = (Y_n, X_n)$ be n i.i.d observations with distribution P that take values in $\{0, 1\} \times \mathbb{R}^p$. Let g be a given classifier and $\alpha_{max} \in (0, 1)$. Consider the penalization function*

$$pen(\alpha) = \frac{1}{(1 - \alpha)} \sqrt{\frac{\ln(n)}{2n}}$$

and define

$$\hat{\alpha} = \arg \min_{\alpha \in [0, \alpha_{max}]} R_{n, \alpha}(g) + pen(\alpha),$$

then the following bound holds,

$$E(R_{\hat{\alpha}}(g)) \leq \inf_{\alpha \in [0, \alpha_{max}]} \left(R_{\alpha}(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{n}(1 - \alpha)} \right) + \frac{1}{(1 - \alpha_{max})} \sqrt{\frac{2\pi}{n}} + \frac{1}{n(1 - \alpha_{max})^2}.$$

This theorem enables to understand the effect of trimming on the classification error. For a given classifier g we fix a maximum level of trimming α_{max} that we do not want to exceed. Then the automatic penalized rule for choosing the trimming level leads to an oracle inequality that warrants that the best classification error is achieved. Similar to model selection rules, the price to pay is a term of order $1/\sqrt{n}$ which does not hamper the classification error. In particular if the classifier g has a small classification error in the sense that $R(g)$ is smaller than some $\alpha < \alpha_{max}$, we achieve to remove the data that are misclassified, leading to a smaller classification error.

A natural extension of this result is the case where we consider a class of classification rules among which the optimal classifier will be selected. A complex class will usually lead us to rules that have a small bias in the sense that they classify well the data in the training sample yet at the expense of larger variance error, usually leading to an overfitting of the classification model. To deal with this necessary control of complexity, the penalties will not only depend on the trimming level as before but also on the complexity of the class of classifiers. This complexity will be measured using the Vapnik-Chervonenkis dimension (VC), see for instant in [14] and references therein. Here \mathcal{F} denotes the set of all classifiers.

Theorem 3.2. *Let ξ_1, \dots, ξ_n be n independent and identically distributed observations with distribution P that take values in $\{0, 1\} \times \mathbb{R}^p$. Let $\{\mathcal{G}_m\}_{m \in \mathbb{N}} \subset \mathcal{F}$ be a family of classes of classifiers with Vapnik-Chervonenkis dimension $V_{\mathcal{G}_m} < \infty$ for all $m \in \mathbb{N}$. Let*

$\alpha_{max} \in (0, 1)$ and let Σ be a non-negative constant. Consider $\{x_m\}_{m \in \mathbb{N}}$ a family of non-negative weights such that

$$\sum_{m \in \mathbb{N}} e^{-x_m} \leq \Sigma < \infty.$$

If we consider the penalization function

$$\text{pen}(\alpha, \mathcal{G}_m) = \sqrt{\frac{\ln(n) + x_m}{2n(1 - \alpha)^2}} + \frac{1}{(1 - \alpha)} \sqrt{\frac{V_{\mathcal{G}_m} \ln(n + 1) + \ln(2)}{n}}$$

and we define

$$(\hat{\alpha}, \hat{m}) = \arg \min_{(\alpha, m) \in [0, \alpha_{max}] \times \mathbb{N}} R_{n, \alpha}(\mathcal{G}_m) + \text{pen}(\alpha, \mathcal{G}_m),$$

the following bound holds

$$\begin{aligned} E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) &\leq \min_{(\alpha, m) \in [0, \alpha_{max}] \times \mathbb{N}} \left(R_{\alpha}(\mathcal{G}_m) + \text{pen}(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n(1 - \alpha)}} \right) \\ &\quad + \frac{1 + \Sigma}{2(1 - \alpha_{max})} \sqrt{\frac{\pi}{2n}} + \frac{1}{n(1 - \alpha_{max})^2}. \end{aligned}$$

Here, again, we obtain a bound similar to the result provided in [20]. The penalty for choosing the trimming parameter depends on the VC dimension of the class of classifiers. Hence, this choice leads to an oracle inequality ensuring the optimality of this selection procedure. As before, the effect of trimming is that it removes an optimal number of data that are misclassified by the collection of classifiers, leading to better classification rates on the set of *good* data.

For a better understanding about the implications of Theorem 3.2 we include next a section which explores this bound for the particular case of linear classifiers.

3.2 Example

Assume we have n i.i.d. observations $(Y_1, X_1), \dots, (Y_n, X_n)$ where $X_i \in \mathbb{R}^p$ and $Y_i \in \{0, 1\}$. We consider the collection of models $\{\mathcal{G}_m\}_{m \in \mathcal{M}}$ where for each m , \mathcal{G}_m is the family of linear classifiers built only using a selection of variables consisting of the first m components of X_i . Set $\mathcal{M} = \{1, \dots, p\}$. For $x \in \mathbb{R}^p$ let $x^{(m)}$ denote the vector consisting of the first m components of x . Define the set of possible classifiers as

$$\mathcal{G}_m = \{g \in \mathcal{F} : g(x) = I_{[a^T x^{(m)} + b \geq 0]}; a \in \mathbb{R}^m; b \in \mathbb{R}\}.$$

Let us also denote by \mathcal{A}_m the collection of all sets

$$\{\{0\} \times \{x : g_m(x) = 1\}\} \cup \{\{1\} \times \{x : g_m(x) = 0\}\}$$

and by \mathcal{B}_m the collection of sets

$$\{x \in \mathbb{R}^p : g_m(x) = 1\}$$

where g_m ranges in \mathcal{G}_m . Using, for instance, Theorem 13.1 and Corollary 13.1 in [10], we have that $V_{\mathcal{G}_m} = V_{\mathcal{A}_m} = m + 1$. Then the penalization function considered in Theorem 3.2 can be written as

$$\text{pen}(\alpha, \mathcal{G}_m) = \sqrt{\frac{\ln(n) + x_m}{2n(1-\alpha)^2}} + \frac{1}{(1-\alpha)} \sqrt{\frac{(m+1)\ln(n+1) + \ln(2)}{n}}.$$

We will choose the family of non-negative weights $x_m = \ln(p)$ for all $m \in \mathcal{M}$ and the universal constant $\Sigma = 1$. If we define

$$\begin{aligned} (\hat{\alpha}, \hat{m}) &= \arg \min_{(\alpha, m) \in [0, \alpha_{\max}] \times \mathcal{M}} \left(R_{n, \alpha}(\mathcal{G}_m) + \sqrt{\frac{\ln(np)}{2n(1-\alpha)^2}} \right. \\ &\quad \left. + \frac{1}{(1-\alpha)} \sqrt{\frac{(m+1)\ln(n+1) + \ln(2)}{n}} \right), \end{aligned}$$

this leads to the following bound

$$\begin{aligned} E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) &\leq \min_{(\alpha, m) \in [0, \alpha_{\max}] \times \mathcal{M}} \left(R_{\alpha}(\mathcal{G}_m) + \sqrt{\frac{\ln(np)}{2n(1-\alpha)^2}} \right. \\ &\quad \left. + \frac{1}{(1-\alpha)} \sqrt{\frac{(m+1)\ln(n+1) + \ln(2)}{n}} + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n(1-\alpha)}} \right) \\ &\quad + \frac{1}{(1-\alpha_{\max})} \sqrt{\frac{\pi}{2n}} + \frac{1}{n(1-\alpha_{\max})^2}. \end{aligned}$$

First we point out that trimming reduces the classification error that may vanish as long as α_{\max} is large enough to remove a sufficient fraction of observations. As in model selection techniques, the last three terms of the right hand side of the inequality are of order $1/\sqrt{n}$ while for a fixed m the third term will be of order $\sqrt{m \ln(n)/n}$. Finally, the second term is of order $\sqrt{\ln(n)/n + \ln(p)/n}$. Hence, as long as $\ln(p)$ is smaller than n , the expected value of the best trimmed classification error for the best class will be small as the number of observations increases.

4 Conclusions and perspectives

In classification theory, many classification rules are affected by the presence of points which are very difficult to classify. When dealing with high dimensional observations or when the number of observations is large, this situation occurs quite often and may drastically hamper the performance of classifiers which take into account all the data. One may be tempted to focus on these points and modify the classification rule to increase their classification ranking for these special points. This is the point of view of boosting algorithms for instance, as described in [12] for example. Yet this is often done at the expense of the complexity of the rule and its ability to be generalized. Hence a practical and maybe pragmatic solution is to consider some of these points as outliers and to simply remove them. Statisticians are reluctant to discard observations, yet in many applications,

in particular when confronted to large amounts of observations, this enables to produce rules that are easier to interpret and that can provide a better understanding of the phenomenon which is studied, provided not too many data are removed from the training sample. This is the typical choice made in several papers but not much is said about the way the outliers are selected and its impact on the classification performance.

This is the reason why we tried to provide in this paper a statistical framework to robust classification by removal of some observations. We provided a method that considers data as outliers based on their classification error by a classifier or a given class of classifiers. Within this framework this procedure enables to select in a data driven way an optimal proportion of observations to be removed in order to achieve a better classification error. The level of trimming and the best classifier are selected simultaneously and we obtain an oracle inequality to assess the quality of this procedure. We think that this result may provide some guidelines to remove outliers for classification problems with theoretical guarantees.

Yet we rely on a minimization of a penalized 0 – 1 loss function which is difficult to handle. A version of this trimming procedure for convex functions that lead to a feasible way of computing the weights is actually under study. We will thus obtain a way of detecting outliers and removing them such that the classification error with this new data set will be theoretically controlled.

5 Appendix

5.1 Technical lemmas

Let $A \subset \{0, 1\} \times \mathbb{R}^p$, we denote $A_i = \{x \in \mathbb{R}^p : (i, x) \in A\}$, for $i = 0, 1$. Obviously $A = (\{0\} \times A_0) \cup (\{1\} \times A_1)$ and the union is disjoint, so for every measurable set $A \subset \{0, 1\} \times \mathbb{R}^p$ and every probability $P \in \{0, 1\} \times \mathbb{R}^p$,

$$P(A) = p_0 P_0(A_0) + p_1 P_1(A_1), \quad (5)$$

where $p_0 = P(\{0\} \times \mathbb{R}^p)$, $p_1 = 1 - p_0$, $P_0(A_0) = P(A|Y = 0) = P(\{0\} \times A_0)/p_0$ and $P_1(A_1) = P(A|Y = 1) = P(\{1\} \times A_1)/p_1$. P_0 and P_1 are probabilities in \mathbb{R}^p . Conversely, from $p_0 \in [0, 1]$ and the probabilities P_0 and P_1 in \mathbb{R}^p the equation (5) defines a probability in $\{0, 1\} \times \mathbb{R}^p$ and the relation is one on one (except for the degenerate cases in which $p_0 = 0$ or $p_0 = 1$), so we can identify the probability P with the object (p_0, P_0, P_1) . We will set $P \equiv (p_0, P_0, P_1)$.

Lemma 5.1. *With the previous notation, if $Q \equiv (q_0, Q_0, Q_1)$ with $q_0 \in (0, 1)$, then $Q \in \mathcal{R}_\alpha(P)$ if and only if*

$$q_0 \leq \frac{p_0}{1 - \alpha}, \quad 1 - q_0 \leq \frac{1 - p_0}{1 - \alpha}, \quad Q_0 \in \mathcal{R}_{1 - \frac{q_0}{p_0}(1 - \alpha)}(P_0) \quad \text{and} \quad Q_1 \in \mathcal{R}_{1 - \frac{1 - q_0}{1 - p_0}(1 - \alpha)}(P_1). \quad (6)$$

Proof. Note first that $q_0 = Q(\{0\} \times \mathbb{R}^p)$, $Q \in \mathcal{R}_\alpha(P)$ implies $q_0 \leq \frac{1}{1 - \alpha} P(\{0\} \times \mathbb{R}^p) = \frac{p_0}{1 - \alpha}$. The same argument shows that $1 - q_0 \leq \frac{1 - p_0}{1 - \alpha}$ if $Q \in \mathcal{R}_\alpha(P)$. Observe that the conditions

$q_0 \leq \frac{p_0}{1-\alpha}$ and $1-q_0 \leq \frac{1-p_0}{1-\alpha}$ guarantee that $0 \leq 1 - \frac{q_0}{p_0}(1-\alpha) \leq 1$ and $0 \leq 1 - \frac{1-q_0}{1-p_0}(1-\alpha) \leq 1$, hence the trimming sets in the statement are well defined. Moreover, if $Q \in \mathcal{R}_\alpha(P)$ then

$$Q_0(A_0) = \frac{Q(\{0\} \times A_0)}{q_0} \leq \frac{1}{(1-\alpha)q_0} P(\{0\} \times A_0) = \frac{1}{(1-\alpha)\frac{q_0}{p_0}} P_0(A_0),$$

which proves that $Q_0 \in \mathcal{R}_{1-\frac{q_0}{p_0}(1-\alpha)}(P_0)$. In a similar way it can be proven that $Q_1 \in \mathcal{R}_{1-\frac{1-q_0}{1-p_0}(1-\alpha)}(P_1)$, which proves that the assumptions (6) are necessary. To prove the sufficiency note that if we have (6) then $q_0 Q_0(A_0) \leq \frac{1}{1-\alpha} P_0(A_0)$, $(1-q_0)Q_1(A_1) \leq \frac{1}{1-\alpha} P_1(A_1)$ and hence

$$Q(A) = q_0 Q_0(A_0) + (1-q_0)Q_1(A_1) \leq \frac{1}{1-\alpha} (p_0 P_0(A_0) + (1-p_0)P_1(A_1)) = \frac{1}{1-\alpha} P(A),$$

which completes the proof. \square

With this identification we now prove the following lemma that will be the first step to prove Proposition 2.1.

Lemma 5.2. *With the previous notation*

$$R_\alpha(g) = \min_{1-\frac{1-p_0}{1-\alpha} \leq q_0 \leq \frac{p_0}{1-\alpha}} \left[\left(q_0 - \frac{p_0}{1-\alpha} P_0(g(x)=0) \right)_+ + \left(1 - q_0 - \frac{1-p_0}{1-\alpha} P_1(g(x)=1) \right)_+ \right]. \quad (7)$$

Proof.

The first step consists in writing the probability Q in terms of (q_0, Q_0, Q_1)

$$\begin{aligned} Q(g(x) \neq y) &= q_0 \int_{(g(x)=1)} \frac{dQ_0}{d\mu} d\mu + (1-q_0) \int_{(g(x)=0)} \frac{dQ_1}{d\mu} d\mu \\ &= \int \left(q_0 I_{(g(x)=1)} \frac{dQ_0}{d\mu} + (1-q_0) I_{(g(x)=0)} \frac{dQ_1}{d\mu} \right) d\mu. \end{aligned} \quad (8)$$

We are looking for the probability that minimizes the probability of error between all the probabilities $Q \in \mathcal{R}_\alpha(P)$, this means we are looking for Q_0 and Q_1 that minimize (8). We are going to make the calculations for Q_0 , Q_1 can be gotten analogously.

As we are minimizing, we are going to concentrate the probability Q_0 in the set $(g(x)=0)$. By Lemma 5.1 we know that $Q_0 \leq \frac{p_0}{q_0(1-\alpha)} P_0$, so the value of Q_0 depends on the value of P_0 . There are two possibilities,

1. $P_0(g(x)=0) \geq \frac{q_0}{p_0}(1-\alpha)$: As $\frac{p_0}{q_0(1-\alpha)} P_0 \geq 1$ we can group all the probability Q_0 in the set $\{x \in \mathbb{R}^p / g(x)=0\}$ and hence $Q_0(g(x)=0) = 1$.
2. $P_0(g(x)=0) < \frac{q_0}{p_0}(1-\alpha)$: Now we can not give to $Q_0(g(x)=0)$ probability 1 because we would be violating the condition in Lemma 5.1, hence $Q_0(g(x)=0) = \frac{P_0(g(x)=0)}{\frac{q_0}{p_0}(1-\alpha)}$.

And in the optimum we will have

$$Q_0(g(x) = 0) = \min \left(\frac{P_0(g(x) = 0)}{\frac{q_0}{p_0}(1 - \alpha)}, 1 \right).$$

As we are concerned in $Q_0(g(x) = 1)$ and Q_0 is a distribution,

$$Q_0(g(x) = 1) = \left(1 - \frac{p_0}{q_0(1 - \alpha)} P_0(g(x) = 0) \right)_+,$$

analogously

$$Q_1(g(x) = 0) = \left(1 - \frac{1 - p_0}{(1 - q_0)(1 - \alpha)} P_1(g(x) = 1) \right)_+.$$

So for a fixed q_0 and Q_0, Q_1 as in Lemma 5.1

$$\begin{aligned} \min_{Q_0, Q_1} Q(g(x) \neq y) &= q_0 \left(1 - \frac{p_0}{q_0(1 - \alpha)} P_0(g(x) = 0) \right)_+ \\ &+ (1 - q_0) \left(1 - \frac{1 - p_0}{(1 - q_0)(1 - \alpha)} P_1(g(x) = 1) \right)_+. \end{aligned}$$

Using that q_0 and $1 - q_0$ are positive lead to (7). The limits for q_0 are obtained from Lemma 5.1. \square

We prove that both the theoretical trimmed error and the empirical error of two close trimming levels is close.

Proposition 5.3. *Let α_1, α_2 be two trimming levels such that $\alpha_2 \in [\alpha_1, \alpha_1 + \frac{1}{n}]$, let α_{max} be such that $\alpha_1 \leq \alpha_2 \leq \alpha_{max} < 1$ and let g be a given classifier, then*

$$R_{\alpha_1}(g) - R_{\alpha_2}(g) \leq \frac{1}{n(1 - \alpha_{max})^2} \quad \text{and} \quad R_{n, \alpha_1}(g) - R_{n, \alpha_2}(g) \leq \frac{1}{n(1 - \alpha_{max})^2}.$$

Proof.

$$\begin{aligned} R_{\alpha_1}(g) - R_{\alpha_2}(g) &= \frac{(R(g) - \alpha_1)_+}{(1 - \alpha_1)} - \frac{(R(g) - \alpha_2)_+}{(1 - \alpha_2)} \\ &= \frac{((1 - \alpha_2)(R(g) - \alpha_1))_+ - ((1 - \alpha_1)(R(g) - \alpha_2))_+}{(1 - \alpha_1)(1 - \alpha_2)} \\ &\leq \frac{1}{(1 - \alpha_1)(1 - \alpha_2)} |R(g) - \alpha_1 - \alpha_2 R(g) + \alpha_1 \alpha_2 - (R(g) - \alpha_2 - \alpha_1 R(g) + \alpha_1 \alpha_2)| \\ &= \frac{1}{(1 - \alpha_1)(1 - \alpha_2)} |-\alpha_1 - \alpha_2 R(g) + \alpha_2 + \alpha_1 R(g)| \\ &= \frac{1}{(1 - \alpha_1)(1 - \alpha_2)} |(R(g) - 1)(\alpha_1 - \alpha_2)| = \frac{1}{(1 - \alpha_1)(1 - \alpha_2)} |R(g) - 1| |\alpha_1 - \alpha_2|. \end{aligned}$$

As we chose α_2 , $|\alpha_1 - \alpha_2| \leq \frac{1}{n}$ and as for every value of α we can bound $\frac{1}{1 - \alpha}$ by $\frac{1}{1 - \alpha_{max}}$ and $|R(g) - 1| \leq 1$, we can conclude that

$$R_{n, \alpha_1}(g) - R_{n, \alpha_2}(g) \leq \frac{1}{n(1 - \alpha_{max})^2}.$$

The proof is identical for the empirical trimmed error. \square

5.2 Proofs

Proof of Proposition 2.1. The result is a direct consequence of the minimization with respect to q_0 of the expression obtained in Lemma 5.2.

First see that $R_\alpha(g) = 0$ if and only if $1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1) \leq \frac{p_0}{1-\alpha}P_0(g(x) = 0)$. Then consider the opposite case.

As we are adding two positive terms, the sum is equal to 0 only if both terms are equal to 0, leading to

$$\left(q_0 - \frac{p_0}{1-\alpha}P_0(g(x) = 0)\right)_+ \leq 0 \Leftrightarrow q_0 \leq \frac{p_0}{1-\alpha}P_0(g(x) = 0),$$

in a similar way we obtain

$$\left(1 - q_0 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1)\right)_+ \leq 0 \Leftrightarrow q_0 \geq 1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1).$$

So $R_\alpha(g) = 0$ if and only if $1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1) \leq \frac{p_0}{1-\alpha}P_0(g(x) = 0)$.

Now consider the case where this inequality does not hold, this means, $1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1) > \frac{p_0}{1-\alpha}P_0(g(x) = 0)$. The first term of (7) is a stepwise lineal function with value 0 until $\frac{p_0}{1-\alpha}P_0(g(x) = 0)$ and increasing with slope 1 since then. The second term is also stepwise linear, in this case it decreases with slope -1 until it reaches 0 in $1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1)$ with value 0 from that point.

Now we are going to see that in this case the interval $[\frac{p_0}{1-\alpha}P_0(g(x) = 0), 1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1)]$ gives us the minimal value of (7). If $1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1) < 1 - \frac{1-p_0}{1-\alpha}$ or $\frac{p_0}{1-\alpha}P_0(g(x) = 0) > \frac{p_0}{1-\alpha}$ we will eliminate non feasible values of q_0 from the optimal set, hence the set of optimal values of q_0 that minimizes $R_\alpha(g)$ are

$$q_0(x) = \begin{cases} [1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1), \frac{p_0}{1-\alpha}P_0(g(x) = 0)] \cap [1 - \frac{1-p_0}{1-\alpha}, \frac{p_0}{1-\alpha}] & \text{si } R(g) \leq \alpha \\ [\frac{p_0}{1-\alpha}P_0(g(x) = 0), 1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1)] \cap [1 - \frac{1-p_0}{1-\alpha}, \frac{p_0}{1-\alpha}] & \text{si } R(g) > \alpha \end{cases} \quad (9)$$

Let us see this. We are going to suppose, for simplicity, that we are in the case

$$1 - \frac{1-p_0}{1-\alpha} \leq \frac{p_0}{1-\alpha}P_0(g(x) = 0) < 1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1) \leq \frac{p_0}{1-\alpha}.$$

Let $I_1 = [1 - \frac{1-p_0}{1-\alpha}, \frac{p_0}{1-\alpha}P_0(g(x) = 0)]$, $I_2 = [\frac{p_0}{1-\alpha}P_0(g(x) = 0), 1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1)]$, $I_3 = [1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1), \frac{p_0}{1-\alpha}]$, we denote

$$R^i = \min_{I_i} \left[\left(q_0 - \frac{p_0}{1-\alpha}P_0(g(x) = 0)\right)_+ + \left(1 - q_0 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1)\right)_+ \right],$$

for $i = 1, 2, 3$. Obviously $R_\alpha(g) = \min R^i$.

In I_1 the first term is 0 because $q_0 \leq \frac{p_0}{1-\alpha}P_0(g(x) = 0)$ and the second term is $1 - \frac{1-p_0}{1-\alpha}P_1(g(x) = 1) - q_0$. As we are looking for a minimization of this value and q_0 is

subtracting, we will give to it the biggest value it can take, that is, the upper bound of the interval. Hence,

$$\begin{aligned} R^1 &= 1 - \frac{1-p_0}{1-\alpha} P_1(g(x)=1) - \frac{p_0}{1-\alpha} P_0(g(x)=0) \\ &= 1 - \frac{(1-p_0)(1-P_1(g(x)=0)) + p_0(1-P_0(g(x)=1))}{1-\alpha} \\ &= 1 - \frac{1-R(g)}{1-\alpha}. \end{aligned}$$

If we are in I_2 none of the terms is going to be 0. First one is $q_0 - \frac{p_0}{1-\alpha} P_0(g(x)=0)$ and second one $1 - \frac{1-p_0}{1-\alpha} P_1(g(x)=1) - q_0$, when we add them, the q_0 in both terms clears and we obtain

$$R^2 = 1 - \frac{p_0}{1-\alpha} P_0(g(x)=0) - \frac{1-p_0}{1-\alpha} P_1(g(x)=1) = 1 - \frac{1-R(g)}{1-\alpha}.$$

Last, in I_3 is the second term which becomes 0, letting the first one as $q_0 - \frac{p_0}{1-\alpha} P_0(g(x)=0)$. In this case q_0 is adding, we want to give the minimum value possible so

$$R^3 = 1 - \frac{1-p_0}{1-\alpha} P_1(g(x)=1) - \frac{p_0}{1-\alpha} P_0(g(x)=0) = 1 - \frac{1-R(g)}{1-\alpha}.$$

And, as we have already said, the minimum is attained at

$$\left[1 - \frac{1-p_0}{1-\alpha} P_1(g(x)=1), \frac{p_0}{1-\alpha} P_0(g(x)=0) \right].$$

Moreover, since $R^1 = R^2 = R^3$, the value of this minimum will be

$$R_\alpha(g) = 1 - \frac{1-R(g)}{1-\alpha}.$$

Putting together both cases we have that $R_\alpha(g)$ reaches its minimum in (9) and, since condition $1 - \frac{1-p_0}{1-\alpha} P_1(g(x)=1) > \frac{p_0}{1-\alpha} P_0(g(x)=0)$ holds if and only if $R(g) > \alpha$, we have that $R_\alpha(g) = 0 \Leftrightarrow R(g) \leq \alpha$ and hence,

$$R_\alpha(g) = \frac{1}{1-\alpha} (R(g) - \alpha)_+.$$

□

Proof of Proposition 2.2. Note that $Err(P) = R(g_B)$ and $Err_\alpha(P) = R_\alpha(g_B^\alpha)$. Recall that

$$\begin{aligned} Err_\alpha(P) &:= \inf_{Q \in \mathcal{R}_\alpha(P)} Err(Q) = \inf_{Q \in \mathcal{R}_\alpha(P)} \inf_g Q(g(x) \neq y) = \inf_g \inf_{Q \in \mathcal{R}_\alpha(P)} Q(g(x) \neq y) \\ &= \inf_g R_\alpha(g) = \min_g \frac{(R(g) - \alpha)_+}{1-\alpha}, \end{aligned}$$

the minimum in the last inequality is due to Proposition 2.1. The infimum is reached so it is a minimum. Moreover we know that this error is minimal when the classifier is Bayes classifier, so

$$Err_\alpha(P) = \frac{(R(g_B) - \alpha)_+}{1-\alpha} = \frac{(Err(P) - \alpha)_+}{1-\alpha}.$$

□

Proof of Proposition 2.4. The first inequality can be proved by

$$\begin{aligned} E(R_{n,\alpha}(g)) &= E\left(\frac{(R_n(g) - \alpha)_+}{1 - \alpha}\right) = \frac{1}{1 - \alpha} E((R_n(g) - \alpha)_+) \\ &\geq \frac{1}{1 - \alpha} (E(R_n(g)) - \alpha)_+ = \frac{1}{1 - \alpha} (R(g) - \alpha)_+ = R_\alpha(g). \end{aligned}$$

Where we have used (4) for the first equality and the property $E(R_n(g)) = R(g)$ and (3) for the two last ones. The inequality comes from applying Jensen inequality, and this is possible due to the fact that $(\cdot)_+$ is a convex function.

For the second inequality we need Proposition 2.1 and by Corollary 2.3,

$$E(R_{n,\alpha}(g)) - R_\alpha(g) = \frac{E((R_n(g) - \alpha)_+) - (R(g) - \alpha)_+}{(1 - \alpha)}. \quad (10)$$

Let X be a random variable such that $X = R_n(g)$, we know from [19] that $E(X) = R(g)$, and let $\varphi(x) = (x - \alpha)_+$. φ is a convex function, so Jensen's inequality can be applied, this means $\varphi(E(X)) \leq E(\varphi(X))$. This function also is 1-Lispchitz and increasing, so it satisfies the property $\varphi(y) - \varphi(x) \leq (y - x)_+$.

As we are not modifying $\frac{1}{1-\alpha}$ we are going to let it aside for the moment and we will focus in the numerator, applying X 's definition, mean's properties and φ 's property mentioned above

$$\begin{aligned} E((R_n(g) - \alpha)_+) - (R(g) - \alpha)_+ &= E(\varphi(X)) - \varphi(E(X)) = E(\varphi(X) - \varphi(E(X))) \\ &\leq E((X - E(X))_+). \end{aligned}$$

Now let Y be a random variable such that $Y \stackrel{d}{=} X$, Y and X are independent, this implies $E(Y) = E_X(Y)$, applying this, again mean's properties, Jensen's inequality (for $(\cdot)_+$) and conditional mean's properties we get

$$\begin{aligned} E((X - E(X))_+) &= E((X - E(Y))_+) = E((X - E_X(Y))_+) = E((E_X(X - Y))_+) \\ &\leq E(E_X((X - Y)_+)) = E((X - Y)_+). \end{aligned}$$

Now we are using that $X - Y$ is a symmetric variable, that it also is a centered variable, variance's property for the sum of two independent variables and that X and Y are identically distributed to obtain

$$\begin{aligned} E((X - Y)_+) &= \frac{1}{2} E(X - Y) \leq \frac{1}{2} (Var(X - Y))^{1/2} = \frac{1}{2} (Var(X) + Var(Y))^{1/2} \\ &= \frac{1}{2} (2Var(X))^{1/2}. \end{aligned}$$

Last we are using variance's properties, the fact that $nX \sim b(n, R(g))$ and that $(1 - R(g)) \leq 1$ and we obtain

$$\frac{1}{2} (2Var(X))^{1/2} = \frac{1}{2} (2 \frac{1}{n^2} Var(nX))^{1/2} = \frac{1}{2} (2 \frac{1}{n^2} nR(g)(1 - R(g)))^{1/2} = \frac{1}{\sqrt{2n}} \sqrt{R(g)}.$$

Joining this with (10), we get

$$E(R_{n,\alpha}(g)) - R_\alpha(g) \leq \frac{\sqrt{R(g)}}{\sqrt{2n(1-\alpha)}}.$$

□

Proof of Theorem 3.1. First consider the case where the trimming parameter only takes value in the discrete set $A = [0, \frac{1}{n}, \dots, \frac{k_0}{n}]$ with $k_0 = \lfloor n\alpha_{max} \rfloor$. By definition $\hat{\alpha}$ satisfies that for all $\alpha \in A$

$$R_{n,\hat{\alpha}}(g) + \text{pen}(\hat{\alpha}) \leq R_{n,\alpha}(g) + \text{pen}(\alpha).$$

This implies that

$$R_{\hat{\alpha}}(g) - R_{\hat{\alpha}}(g) + R_{n,\hat{\alpha}}(g) + \text{pen}(\hat{\alpha}) \leq R_\alpha(g) - R_\alpha(g) + R_{n,\alpha}(g) + \text{pen}(\alpha)$$

or, what is the same

$$R_{\hat{\alpha}}(g) \leq R_\alpha(g) + \text{pen}(\alpha) + (R_{n,\alpha}(g) - R_\alpha(g)) - \text{pen}(\hat{\alpha}) + (R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g)). \quad (11)$$

Let us focus in the inside of the parenthesis,

$$R_{n,\alpha}(g) - R_\alpha(g) = [R_{n,\alpha}(g) - E(R_{n,\alpha}(g))] + [E(R_{n,\alpha}(g)) - R_\alpha(g)],$$

by Proposition 2.4 the second brace can be bounded by $\frac{\sqrt{R(g)}}{\sqrt{2n(1-\alpha)}}$. For first brace we will apply McDiarmid's inequality taking $R_{n,\alpha}(g) = F(\xi_1, \dots, \xi_n)$ where $\xi_i = (Y_i, X_i)$. As

$$|F(\xi_1, \dots, \xi_i, \dots, \xi_n) - F(\xi_1, \dots, \xi'_i, \dots, \xi_n)| \leq \frac{1}{n(1-\alpha)},$$

we can apply the inequality and hence

$$P(R_{n,\alpha}(g) - E(R_{n,\alpha}(g)) \geq t) \leq e^{-2t^2n(1-\alpha)^2}.$$

Given $z > 0$ take $t = \sqrt{\frac{z}{2n(1-\alpha)^2}}$, we get

$$P\left(R_{n,\alpha}(g) - E(R_{n,\alpha}(g)) \geq \sqrt{\frac{z}{2n(1-\alpha)^2}}\right) \leq e^{-z}.$$

Joining this with (11), we get that, except in a set of probability not greater than e^{-z} ,

$$R_{\hat{\alpha}}(g) \leq R_\alpha(g) + \text{pen}(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{2n(1-\alpha)}} + \sqrt{\frac{z}{2n(1-\alpha)^2}} - \text{pen}(\hat{\alpha}) + (R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g)). \quad (12)$$

We are going to focus now in the other parenthesis. As we saw in Proposition 2.4

$$R_{\hat{\alpha}}(g) - R_{n,\hat{\alpha}}(g) \leq \sup_{\alpha \in A} (R_\alpha(g) - R_{n,\alpha}(g)) \leq \sup_{\alpha \in A} (E(R_{n,\alpha}(g)) - R_{n,\alpha}(g)).$$

Applying again McDiarmid's inequality taking this time $t = \sqrt{\frac{\ln(n)+z}{2n(1-\alpha)^2}}$ we have $\forall \alpha' \in A$

$$P \left(E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) \geq \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}} \right) \leq \frac{1}{n} e^{-z}.$$

As we were interested in calculating this probability for $\hat{\alpha}$ we have

$$\begin{aligned} & P \left(\sup_{\alpha \in A} (E(R_{n,\alpha}(g)) - R_{n,\alpha}(g)) \geq \sqrt{\frac{\ln(n)+z}{2n(1-\hat{\alpha})^2}} \right) \\ & \leq \sum_{\alpha' \in A} P \left(E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) \geq \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}} \right) \\ & \leq n \frac{1}{n} e^{-z} \leq e^{-z}. \end{aligned}$$

Hence with probability at least $1 - e^{-z}$

$$E(R_{n,\hat{\alpha}}(g)) - R_{n,\hat{\alpha}}(g) \leq \sqrt{\frac{\ln(n)+z}{2n(1-\hat{\alpha})^2}}. \quad (13)$$

Now let us consider the complete interval. If $\alpha' \in [0, \alpha_{max}]$ there exists $\alpha'' \in A$ such that $\alpha'' \leq \alpha' \leq \alpha'' + \frac{1}{n}$. Then by Proposition 5.3, in the set where (13) is satisfied we have

$$\begin{aligned} & E(R_{n,\alpha'}(g)) - R_{n,\alpha'}(g) \\ & = E(R_{n,\alpha''}(g)) - R_{n,\alpha''}(g) + E(R_{n,\alpha'}(g) - R_{n,\alpha''}(g)) - (R_{n,\alpha'}(g) - R_{n,\alpha''}(g)) \\ & \leq \sqrt{\frac{\ln(n)+z}{2n(1-\alpha'')^2}} + \frac{1}{n(1-\alpha_{max})^2} \leq \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}} + \frac{1}{n(1-\alpha_{max})^2} \\ & \leq \sqrt{\frac{\ln(n)}{2n(1-\alpha')^2}} + \sqrt{\frac{z}{2n(1-\alpha')^2}} + \frac{1}{n(1-\alpha_{max})^2} \end{aligned}$$

for all $\alpha' \in [0, \alpha_{max}]$.

If we take as penalty

$$pen(\alpha) = \sqrt{\frac{\ln(n)}{2n(1-\alpha)^2}}$$

and we substitute in (12). Given that, by Proposition 2.4, $R_{\alpha'}(g) \leq E(R_{n,\alpha'}(g))$, with probability at least $1 - 2e^{-z}$

$R_{\hat{\alpha}}(g)$

$$\begin{aligned} & \leq R_{\alpha}(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{2n(1-\alpha)}} + \sqrt{\frac{z}{2n(1-\alpha)^2}} - pen(\hat{\alpha}) + \sqrt{\frac{\ln(n)+z}{2n(1-\alpha')^2}} + \frac{1}{n(1-\alpha_{max})^2} \\ & \leq R_{\alpha}(g) + pen(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{2n(1-\alpha)}} + 2\sqrt{\frac{z}{2n(1-\alpha_{max})^2}} + \frac{1}{n(1-\alpha_{max})^2}. \end{aligned}$$

Integrating with respect to z and taking the infimum for $\alpha \in [0, \alpha_{max}]$ we can conclude that

$$E(R_{\hat{\alpha}}(g)) \leq \inf_{\alpha \in [0, \alpha_{max}]} \left(R_{\alpha}(g) + \text{pen}(\alpha) + \frac{\sqrt{R(g)}}{\sqrt{n}(1-\alpha)} \right) + \frac{1}{(1-\alpha_{max})} \sqrt{\frac{2\pi}{n}} + \frac{1}{n(1-\alpha_{max})^2}.$$

□

Proof of Theorem 3.2. To proof the theorem we need the following elemental result whose proof is omitted.

Lemma 5.4. *Given two functions f and g and a real number $k > 0$,*

$$|f(x) - g(x)| \leq k \Rightarrow \left| \sup_x f(x) - \sup_x g(x) \right| \leq k,$$

$$|f(x) - g(x)| \leq k \Rightarrow \left| \min_x f(x) - \min_x g(x) \right| \leq k.$$

As in the previous Theorem we define the set $A = \{0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{k_0}{n}\}$ with $k_0 = \lfloor n\alpha_{max} \rfloor$. Then, by definition, $\hat{\alpha}$ and \hat{m} satisfy that for all $\alpha \in A$ and $m \in \mathbb{N}$

$$R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}}) + \text{pen}(\hat{\alpha}, \mathcal{G}_{\hat{m}}) \leq R_{n,\alpha}(\mathcal{G}_m) + \text{pen}(\alpha, \mathcal{G}_m).$$

Adding and subtracting $R_{\hat{\alpha}}(\mathcal{G}_m)$ and $R_{\alpha}(\mathcal{G}_m)$ and organizing the terms we get the following inequality. We bound the remaining terms in the parenthesis,

$$R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) \leq R_{\alpha}(\mathcal{G}_m) + \text{pen}(\alpha, \mathcal{G}_m) + (R_{n,\alpha}(\mathcal{G}_m) - R_{\alpha}(\mathcal{G}_m)) - \text{pen}(\hat{\alpha}, \mathcal{G}_{\hat{m}}) + (R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}})).$$

First we are going to bound

$$R_{n,\alpha}(\mathcal{G}_m) - R_{\alpha}(\mathcal{G}_m) = \min_{g \in \mathcal{G}_m} R_{n,\alpha}(\mathcal{G}_m) - \min_{g \in \mathcal{G}_m} R_{\alpha}(\mathcal{G}_m) \leq R_{n,\alpha}(g') - R_{\alpha}(g')$$

with $g' := \arg \min_{g \in \mathcal{G}_m} R_{\alpha}(g)$. We are now in the same conditions as in Theorem 3.1 and we can bound these quantities except on a set of probability not greater than e^{-z} , with a given $z > 0$ by

$$R_{n,\alpha}(g') - R_{\alpha}(g') \leq \frac{R(g')}{\sqrt{2n}(1-\alpha)} + \sqrt{\frac{z}{2n(1-\alpha)^2}},$$

which leads us to

$$R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) \leq R_{\alpha}(\mathcal{G}_m) + \text{pen}(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(g')}}{\sqrt{2n}(1-\alpha)} + \sqrt{\frac{z}{2n(1-\alpha)^2}} - \text{pen}(\hat{\alpha}, \mathcal{G}_{\hat{m}}) + (R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}})). \quad (14)$$

Now we want to bound

$$R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}}) \leq \sup_{(\alpha', m') \in A \times \mathbb{N}} (R_{\alpha'}(\mathcal{G}_{m'}) - R_{n,\alpha'}(\mathcal{G}_{m'})) \leq \sup_{(\alpha', m') \in A \times \mathbb{N}} \sup_{g \in \mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g)).$$

Let us focus on

$$\sup_{g \in \mathcal{G}_m} (R_\alpha(g) - R_{n,\alpha}(g)) = E \left(\sup_{g \in \mathcal{G}_m} (R_\alpha(g) - R_{n,\alpha}(g)) \right) \quad (15)$$

$$+ \left[\sup_{g \in \mathcal{G}_m} (R_\alpha(g) - R_{n,\alpha}(g)) - E \left(\sup_{g \in \mathcal{G}_m} (R_\alpha(g) - R_{n,\alpha}(g)) \right) \right] \quad (16)$$

To bounding (16) we will make use again of McDiarmid's inequality. First we need to see that the bounded difference conditions is met.

We define $Z := f(\xi_1, \dots, \xi_n) = \sup_{g \in \mathcal{G}_m} (R_\alpha(g) - R_{n,\alpha}(g))$ and $Z^{(i)} := f(\xi_1, \dots, \xi'_i, \dots, \xi_n)$, we want to prove

$$|Z - Z^{(i)}| \leq c_i, \quad (17)$$

for certain constants c_i . The empirical error $R_{n,\alpha}(g)$ is defined as in (1) and $R_{n,\alpha}^{(i)}(g)$ as the empirical error associated to the sample $\xi_1, \dots, \xi'_i, \dots, \xi_n$. We start from

$$|(R_\alpha(g) - R_{n,\alpha}(g)) - (R_\alpha(g) - R_{n,\alpha}^{(i)}(g))|$$

which implies, using Lemma 5.4, that (17).

$$|R_{n,\alpha}(g) - R_{n,\alpha}^{(i)}(g)| = \left| \min_{(w_1, \dots, w_n)} \sum_j w_j I_{(g(X_j) \neq Y_j)} - \min_{(w_1, \dots, w_n)} \sum_j w_j I_{(g(X'_j) \neq Y'_j)} \right|,$$

where (Y', X') stands for the sample $\xi_1, \dots, \xi'_i, \dots, \xi_n$. For a vector (w_1, \dots, w_n) that satisfies the conditions (2),

$$\left| \sum_j w_j I_{(g(X_j) \neq Y_j)} - \sum_j w_j I_{(g(X'_j) \neq Y'_j)} \right| = w_j |I_{(g(X_i) \neq Y_i)} - I_{(g(X'_i) \neq Y'_i)}| \leq \frac{1}{n(1-\alpha)}.$$

And using the second statement of Lemma 5.4 leads to

$$|R_{n,\alpha}(g) - R_{n,\alpha}^{(i)}(g)| \leq \frac{1}{n(1-\alpha)},$$

or written in a different way

$$|(R_\alpha(g) - R_{n,\alpha}(g)) - (R_\alpha(g) - R_{n,\alpha}^{(i)}(g))| \leq \frac{1}{n(1-\alpha)}.$$

Applying again Lemma 5.4, we get to (17) with $c_i = \frac{1}{n(1-\alpha)}$. Now we can use McDiarmid's inequality to prove

$$P \left(\sup_{g \in \mathcal{G}_m} (R_\alpha(g) - R_{n,\alpha}(g)) - E \left(\sup_{g \in \mathcal{G}_m} (R_\alpha(g) - R_{n,\alpha}(g)) \right) \geq \sqrt{\frac{\ln(n) + z + x_m}{2n(1-\alpha)^2}} \right) \leq \frac{1}{n} e^{-z-x_m}. \quad (18)$$

To bound (15) we will use Vapnik-Chervonenkis theory from [11] or [19]. Before we are able to apply this theory we need to transform our functions in suitable functions. For

this we will use the equalities (3) and (4) and the fact that the function positive part, defines as $X_+ := \max(0, X)$, is Lipschitz.

$$\begin{aligned}
E \left(\sup_{g \in \mathcal{G}_m} (R_\alpha(g) - R_{n,\alpha}(g)) \right) &= \frac{1}{1-\alpha} E \left(\sup_{g \in \mathcal{G}_m} ((R(g) - \alpha)_+(R_n(g) - \alpha)_+) \right) \\
&\leq \frac{1}{1-\alpha} E \left(\sup_{g \in \mathcal{G}_m} |R(g) - R_n(g)| \right) \\
&\leq \frac{2}{1-\alpha} \sqrt{\frac{V_{\mathcal{A}_{\mathcal{G}_m}} \ln(n+1) + \ln(2)}{n}}. \tag{19}
\end{aligned}$$

The last inequality comes from section 4.2 in [11]. Joining (18) and (19) we get $\forall \alpha' \in A$ and $\forall m' \in \mathbb{N}$

$$P \left(\sup_{g \in \mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g)) \geq \sqrt{\frac{\ln(n) + z + x_{m'}}{2n(1-\alpha')^2}} + \frac{2}{1-\alpha'} \sqrt{\frac{V_{\mathcal{G}_{m'}} \ln(n+1) + \ln(2)}{n}} \right) \leq \frac{1}{n} e^{-z-x_{m'}}. \tag{20}$$

As we are looking for a bound for $R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}})$, we have that

$$\begin{aligned}
P \left(\bigcup_{(\alpha', m') \in A \times \mathbb{N}} \sup_{g \in \mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g)) \geq \sqrt{\frac{\ln(n) + z + x_{m'}}{2n(1-\alpha')^2}} + \frac{2}{1-\alpha'} \sqrt{\frac{V_{\mathcal{G}_{m'}} \ln(n+1) + \ln(2)}{n}} \right) \\
\leq \sum_{\alpha' \in A} \sum_{m' \in \mathbb{N}} P \left(R_{\alpha'}(g) - R_{n,\alpha'}(g) \geq \sqrt{\frac{\ln(n) + z + x_{m'}}{2n(1-\alpha')^2}} + \frac{2}{1-\alpha'} \sqrt{\frac{V_{\mathcal{G}_{m'}} \ln(n+1) + \ln(2)}{n}} \right) \\
\leq \sum_{\alpha' \in A} \sum_{m' \in \mathbb{N}} \frac{1}{n} e^{-z-x_{m'}} \leq \sum_{m' \in \mathbb{N}} e^{-z-x_{m'}} \leq \Sigma e^{-z}.
\end{aligned}$$

Considering now the complete interval, if $\alpha' \in [0, \alpha_{max}]$, then $\exists \alpha'' \in A$ such that $\alpha'' \leq \alpha' \leq \alpha'' + \frac{1}{n}$. So from (20), with probability greater than $\frac{1}{n} e^{-z-x_{m'}}$,

$$\sup_{g \in \mathcal{G}_{m'}} (R_{\alpha''}(g) - R_{n,\alpha''}(g)) \leq \sqrt{\frac{\ln(n) + z + x_{m'}}{2n(1-\alpha'')^2}} + \frac{2}{1-\alpha''} \sqrt{\frac{V_{\mathcal{G}_{m'}} \ln(n+1) + \ln(2)}{n}},$$

then for all $\alpha' \in [0, \alpha_{max}]$

$$\begin{aligned}
&\sup_{g \in \mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g)) \\
&= \sup_{g \in \mathcal{G}_{m'}} (R_{\alpha'}(g) - R_{n,\alpha'}(g) + R_{\alpha''}(g) - R_{\alpha''}(g) + R_{n,\alpha''}(g) - R_{n,\alpha''}(g)) \\
&= \sup_{g \in \mathcal{G}_{m'}} ([R_{\alpha''}(g) - R_{n,\alpha''}(g)] + [R_{\alpha'}(g) - R_{\alpha''}(g)] + [R_{n,\alpha''}(g) - R_{n,\alpha'}(g)]) \\
&\leq \sqrt{\frac{\ln(n) + z + x_{m'}}{2n(1-\alpha'')^2}} + \frac{2}{1-\alpha''} \sqrt{\frac{V_{\mathcal{G}_{m'}} \ln(n+1) + \ln(2)}{n}} + \frac{1}{n(1-\alpha_{max})^2} \\
&\leq \sqrt{\frac{\ln(n) + x_{m'}}{2n(1-\alpha')^2}} + \sqrt{\frac{z}{2n(1-\alpha')^2}} + \frac{2}{1-\alpha'} \sqrt{\frac{V_{\mathcal{G}_{m'}} \ln(n+1) + \ln(2)}{n}} + \frac{1}{n(1-\alpha_{max})^2}.
\end{aligned}$$

Where the next-to-last inequality comes from applying Proposition 5.3 and that $R_{n,\alpha'}(g) \leq R_{n,\alpha''}(g)$ and hence $R_{n,\alpha'}(g) - R_{n,\alpha''}(g) \leq 0$ and the last one comes from $\alpha'' \leq \alpha'$. We can conclude that

$$R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) - R_{n,\hat{\alpha}}(\mathcal{G}_{\hat{m}}) \leq \sqrt{\frac{\ln(n) + z + x_{\hat{m}}}{2n(1 - \hat{\alpha})^2}} + \frac{2}{1 - \hat{\alpha}} \sqrt{\frac{V_{\mathcal{G}_{\hat{m}}} \ln(n+1) + \ln(2)}{n}} + \frac{1}{n(1 - \alpha_{max})^2}.$$

Going back to (14), except in a set of probability not greater than $(\Sigma + 1)e^{-z}$

$$\begin{aligned} R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq R_{\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n(1 - \alpha)}} + \sqrt{\frac{z}{2n(1 - \alpha)^2}} - pen(\hat{\alpha}, \mathcal{G}_{\hat{m}}) \\ &+ \sqrt{\frac{\ln(n) + x_{\hat{m}}}{2n(1 - \hat{\alpha})^2}} + \sqrt{\frac{z}{2n(1 - \hat{\alpha})^2}} + \frac{2}{1 - \hat{\alpha}} \sqrt{\frac{V_{\mathcal{G}_{\hat{m}}} \ln(n+1) + \ln(2)}{n}} + \frac{1}{n(1 - \alpha_{max})^2}. \end{aligned}$$

Considering

$$pen(\alpha, \mathcal{G}_m) = \sqrt{\frac{\ln(n+1) + x_m}{2n(1 - \alpha)^2}} + \frac{1}{(1 - \alpha)} \sqrt{\frac{V_{\mathcal{G}_m} \ln(n+1) + \ln(2)}{n}},$$

we have

$$\begin{aligned} R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}}) &\leq R_{\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n(1 - \alpha)}} + \sqrt{\frac{z}{2n(1 - \alpha)^2}} + \sqrt{\frac{z}{2n(1 - \hat{\alpha})^2}} + \frac{1}{n(1 - \alpha_{max})^2} \\ &\leq R_{\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n(1 - \alpha)}} + \sqrt{\frac{2z}{n(1 - \frac{k_0}{n})^2}} + \frac{1}{n(1 - \alpha_{max})^2}. \end{aligned}$$

Now grouping and integrating with respect to z ,

$$\begin{aligned} E(R_{\hat{\alpha}}(\mathcal{G}_{\hat{m}})) &\leq \min_{(\alpha, m) \in [0, \alpha_{max}] \times \mathbb{N}} \left(R_{\alpha}(\mathcal{G}_m) + pen(\alpha, \mathcal{G}_m) + \frac{\sqrt{R(\mathcal{G}_m)}}{\sqrt{2n(1 - \alpha)}} \right) \\ &+ \frac{1 + \Sigma}{2(1 - \alpha_{max})} \sqrt{\frac{\pi}{2n}} + \frac{1}{n(1 - \alpha_{max})^2}. \end{aligned}$$

□

References

- [1] Alfons, A., Croux, C. and Gelper, S. (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics* **7**(1), 226-248.
- [2] Álvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J. A. and Matrán, C. (2012). Similarity of samples and trimming. *Bernoulli* **18**(2), 606-634.
- [3] Álvarez-Esteban, P. C., del Barrio, E., Cuesta-Albertos, J. A. and Matrán, C. (2012). Trimmed comparison of distributions. *Journal of the American Statistical Association* **47**(2), 358-375.

- [4] Bartlett, P. L. and Wegkamp, M. H. (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research* **9**(Aug), 1823-1840.
- [5] Chen, X., Wang, Z. J. and McKeown, M. J. (2010). Asymptotic analysis of robust LASSOs in the presence of noise with large variance. *IEEE Transactions on Information Theory* **56**(10), 5131-5149.
- [6] Christmann, A. and Steinwart, I. (2004). On robustness properties of convex risk minimization methods for pattern recognition. *J. Mach. Learn. Res.* **5**(Aug), 1007-1034.
- [7] Cristianini, N. and Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.
- [8] Cuesta-Albertos, J. A., Gordaliza, A. and Matrán, C. (1997). Trimmed k -means: an attempt to robustify quantizers. *The Annals of Statistics* **25**(2), 553-576.
- [9] Debruyne, M. (2009). An outlier map for support vector machine classification. *The Annals of Applied Statistics* **3**(4), 1566-1580.
- [10] Devroye, L., Györfi, L. and Lugosi, G. (2013). A probabilistic theory of pattern recognition. Springer Science & Business Media.
- [11] Devroye, L. and Lugosi, G. (2012). Combinatorial methods in density estimation. Springer Science & Business Media.
- [12] Freund, Y. and Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. *European conference on computational learning theory*, 23-37. Springer.
- [13] García-Escudero, L. A., Gordaliza, A. and Matrán, C. (2003). Trimming tools in exploratory data analysis. *Journal of Computational and Graphical Statistics* **12**(2), 434-449.
- [14] Van De Geer, S. (2000). *Empirical Processes in M-estimation* 6.
- [15] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). Robust statistics. John Wiley & Sons, Inc., New York.
- [16] Hastie, T., Tibshirani, R. and Friedman, J. (2009). The elements of statistical learning (Vol 1). Springer, Berlin: Springer series in statistics.
- [17] Herbei, R. and Wegkamp, M. H. (2006). Classification with reject option. *Canadian Journal of Statistics* **34**(4), 709-721. Wiley Online Library.
- [18] Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35**, 73-101.
- [19] Lugosi, G. (2002). Pattern classification and learning theory. *Principles of nonparametric learning* 1-56. Springer Vienna.

- [20] Massart, P. (2007). Concentration inequalities and model selection. Springer Berlin.
- [21] Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American statistical association* **79**(388), 871-880.
- [22] Maronna, R. A. (2011). Robust ridge regression for high-dimensional data. *Technometrics* **53**(1), 44-53. Taylor & Francis.